

QUASI-NEWTON METHODS: SUPERLINEAR CONVERGENCE WITHOUT LINE SEARCH FOR SELF-CONCORDANT FUNCTIONS

WENBO GAO AND DONALD GOLDFARB[†]

Industrial Engineering and Operations Research, Columbia University

ABSTRACT. We derive a *curvature-adaptive* step size for gradient-based iterative methods, including quasi-Newton methods, for minimizing self-concordant functions. This step size has a simple expression that can be computed analytically; hence, line searches are not needed. We show that using this step size in the BFGS method (and quasi-Newton methods generally) results in superlinear convergence for strongly convex self-concordant functions. We present numerical experiments comparing gradient descent and BFGS methods using the curvature-adaptive step size to traditional methods on logistic regression problems.

1. INTRODUCTION

Iterative optimization algorithms produce a sequence of points converging to the optimal solution. At each step, the algorithm must select both a *direction* d_k and a *step size* t_k , taking a step $s_k = t_k d_k$. Choosing an appropriate step size t_k is necessary for the algorithm to perform efficiently, both in theory and in practice.

Theoretical proofs of global convergence generally assume one of the following approaches for selecting the step sizes:

- (1) The step sizes are obtained from line search
- (2) The step size is a ‘sufficiently small’ constant

Both of these approaches have significant disadvantages in practice. Line search can be costly to perform, and is often prohibitively costly for many common objective functions arising from empirical risk minimization. In contrast, constant step sizes $t_k = t$ make computing the step $t d_k$ easy, but determining an appropriate constant t is difficult. The value of t required in the theoretical analysis is often too small for practical purposes, and moreover, is impossible to compute without knowledge of unknown parameters (the Lipschitz constant of ∇f). A single constant step size may also be highly suboptimal, as the iterates transition between regions with different curvature.

The basic idea for a step size determined by the local curvature was developed by Nesterov, who introduced the *damped Newton method* [11]. This idea is closely related to a well-behaved class of functions known as *self-concordant functions*, which we define in Section 3. When applied to a self-concordant function f , the damped Newton method is globally convergent and locally converges quadratically. These results were greatly extended in recent work.

E-mail address: wg2279@columbia.edu, goldfarb@columbia.edu.

Date: December 20, 2016.

2010 *Mathematics Subject Classification.* 90C53, 90C30.

[†]Research of this author was supported in part by NSF Grant CCF-1527809.

- (1) Tran-Dinh et al. [18] propose a variable-metric framework for composite self-concordant minimization, which includes proximal damped Newton, proximal quasi-Newton, and proximal gradient descent. They establish that proximal damped Newton is globally convergent and locally quadratically convergent, and that proximal gradient descent is globally convergent and locally linearly convergent.
- (2) Zhang and Xiao [19] propose a distributed method for self-concordant empirical loss functions, based on the damped Newton method, and establish its convergence.
- (3) Lu [10] proposes a randomized block proximal damped Newton method for composite self-concordant minimization, and establishes its convergence.

We extend the theory of self-concordant minimization developed by Nesterov and Nemirovski [12] by proposing a variable-metric framework, similar to that of Tran-Dinh et al. [18], and derive a step size that is optimal with respect to an upper bound on the decrease in the objective value. We first prove that scaled gradient methods that use this step size are globally R -linearly convergent on strongly convex self-concordant functions. We then prove that the BFGS method, using this step size, is globally convergent for functions that are self-concordant, bounded below, and have bounded Hessian, and furthermore, is Q -superlinearly convergent when the function is strongly convex and self-concordant.

Our paper is organized as follows. In Section 2, we discuss the notation and assumptions used throughout. In Section 3, we define the class of self-concordant functions and describe their essential properties. In Section 4, we introduce our variable-metric framework for self-concordant minimization and derive what we call the *curvature-adaptive* step size. In Section 5, we apply our approach to scaled gradient methods, and give a simple proof that these methods are globally R -linearly convergent on strongly convex self-concordant functions. In Section 6, we analyze the convergence of the BFGS method with curvature-adapted step sizes. In Section 7, we present numerical experiments testing our new methods on logistic regression problems. In Section 8, we discuss stochastic extensions of adaptive methods and directions for further research.

2. PRELIMINARIES

We use $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to denote the objective function, and $g(\cdot), G(\cdot)$ denote the gradient and Hessian of f respectively. In the context of a sequence of points $\{x_k\}_{k=0}^\infty$, we write g_k for $g(x_k)$ and G_k for $G(x_k)$. Unless stated otherwise, the function f is assumed to be three-times continuously differentiable and self-concordant.

The norm $\|\cdot\|$ denotes the 2-norm, and when applied to a matrix, the operator 2-norm.

3. SELF-CONCORDANT FUNCTIONS

The notion of *self-concordant* functions was first introduced by Nesterov and Nemirovski [12] for their analysis of Newton's method in the context of interior-point methods. Nesterov [11] provides a clear exposition and motivates self-concordancy by observing that, while Newton's method is invariant under affine transformations, the convergence analysis makes use of norms which are *not* invariant. To remedy this, Nesterov and Nemirovski replace the Euclidean norm by an invariant local norm, and replace the assumption of Lipschitz continuity of the Hessian $G(x)$ by the self-concordancy of f .

Definition. Let f be a convex function. The local norm of $h \in \mathbb{R}^n$ at the point x is given by

$$\|h\|_x = \sqrt{h^T G(x) h}$$

Definition. A convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is self-concordant if there exists a constant κ such that for every $x \in \mathbb{R}^n$ and every $h \in \mathbb{R}^n$, we have

$$|\nabla^3 f(x)[h, h, h]| \leq \kappa (\nabla^2 f(x)[h, h])^{3/2}$$

If $\kappa = 2$, f is standard self-concordant. Any self-concordant function can be scaled to be standard self-concordant; the scaled function $\frac{1}{4}\kappa^2 f$ is standard self-concordant. Hence, we assume all self-concordant functions have $\kappa = 2$, unless stated otherwise.

There is also an equivalent definition which is frequently useful.

Theorem 3.1 (Lemma 4.1.2, [11]). A convex function f is self-concordant if and only if for every $x \in \mathbb{R}^n$ and all $u_1, u_2, u_3 \in \mathbb{R}^n$, we have

$$|\nabla^3 f(x)[u_1, u_2, u_3]| \leq 2 \prod_{i=1}^3 \|u_i\|_x$$

The next inequalities are fundamental for self-concordant functions. These results are well known (see §4.1.4 of [11]), but for completeness, we provide a proof.

Lemma 3.2. Let f be self-concordant, and let $x, h \in \mathbb{R}^n$. Let $r = \|h\|_x$. Then for all $t \geq 0$,

$$(3.1) \quad f(x + th) \geq f(x) + tg(x)^T h + rt - \log(1 + rt)$$

and

$$(3.2) \quad g(x + th)^T h \geq g(x)^T h + \frac{r^2 t}{1 + rt}$$

For all $0 \leq t < \frac{1}{r}$,

$$(3.3) \quad f(x + th) \leq f(x) + tg(x)^T h - rt - \log(1 - rt)$$

and

$$(3.4) \quad g(x + th)^T h \leq g(x)^T h + \frac{r^2 t}{1 - rt}$$

Proof. Define $\phi : \mathbb{R} \rightarrow \mathbb{R}$ by $\phi(t) = h^T \nabla^2 f(x + th)h$. Since f is three-times continuously differentiable, $\phi(t)$ is continuously differentiable and its derivative satisfies

$$|\phi'(t)| = |\nabla^3 f(x + th)[h, h, h]| \leq 2(\nabla^2 f(x + th)[h, h])^{3/2} = 2\phi(t)^{3/2}$$

Therefore, by the definition of self-concordancy,

$$\left| \frac{d}{dt} \phi(t)^{-1/2} \right| = \frac{1}{2} |\phi(t)^{-3/2} \phi'(t)| \leq 1$$

Define $\psi(s) = \frac{d}{dt} \phi(t)^{-1/2} \big|_{t=s}$, so the above equation is equivalent to $|\psi(s)| \leq 1$. By Taylor's Theorem, there exists a point $u \in (0, t)$ such that $\phi(t)^{-1/2} = \phi(0)^{-1/2} + t\psi(u)$. We deduce that

$$\phi(0)^{-1/2} - t \leq \phi(t)^{-1/2} \leq \phi(0)^{-1/2} + t$$

Note that $r = \phi(0)^{1/2}$. Rearranging the upper bound, we find that for all $t \geq 0$,

$$(3.5) \quad \phi(t) \geq \frac{r^2}{(1 + rt)^2}$$

Similarly, we find that for $0 \leq t < \frac{1}{r}$,

$$(3.6) \quad \phi(t) \leq \frac{r^2}{(1 - rt)^2}$$

Integrating (3.5) yields the inequalities (3.1), (3.2), and integrating (3.6) produces (3.3), (3.4). \square

4. CURVATURE-ADAPTIVE STEP SIZES

We define a general framework for an iterative method with step sizes determined by the local curvature. At each step, we compute a descent direction $d_k = -H_k g_k$, where H_k is a positive definite matrix, and a step size t_k given by $t_k = \frac{\alpha_k}{1 + \alpha_k \delta_k}$, where $\delta_k = \|d_k\|_{x_k}$ and α_k will be determined later. Furthermore, let us define $s_k = t_k d_k$ and $\rho_k = g_k^T H_k g_k$. We then advance to the point $x_{k+1} = x_k + s_k$.

We will refer to the step size t_k as the *curvature-adaptive* step size, or simply the *adaptive* step size. A method within our framework will be referred to as an *adaptive* method.

Note that this framework encompasses several classical methods. When $H_k = I$ for all k , the resulting method is gradient descent. When $H_k = (\nabla^2 f(x_k))^{-1}$, we recover the *damped Newton method* proposed by Nesterov. When H_k is an approximation of $(\nabla^2 f(x_k))^{-1}$, we obtain quasi-Newton methods. In particular, we will focus on the case where H_k evolves according to the BFGS update formula.

Using the results of Section 3, we can determine the optimal choice of step size t_k .

Lemma 4.1. *Let f be self-concordant. Suppose that we take a step $s_k = t_k d_k$, where $d_k = -H_k g_k$ and $t_k = \frac{\alpha_k}{1 + \alpha_k \delta_k}$. The upper bound on the decrease in objective value implied by inequality (3.1) is obtained by $\alpha_k = \frac{\rho_k}{\delta_k^2}$. For this step size, (3.1) can be expressed as*

$$f(x_{k+1}) \leq f(x_k) - \omega(\eta_k)$$

where $\eta_k = \frac{\rho_k}{\delta_k}$ and $\omega : \mathbb{R} \rightarrow \mathbb{R}$ is the function $\omega(z) = z - \log(1 + z)$.

Proof. We fix the index k and omit the indices for brevity. Apply inequality (3.1) with $h = d$ and $t = t_k$. Observe that the required conditions hold, as $r = \delta$ and

$$t = \frac{\alpha}{1 + \alpha \delta} < \frac{1}{r}$$

Therefore, inequality (3.1) implies that

$$f(x + s) \leq f(x) + t g^T h - \delta t - \log(1 - \delta t)$$

Using our definition $\rho = g^T H g$, we have

$$f(x + s) \leq f(x) - \frac{\alpha(\rho + \delta)}{1 + \alpha \delta} + \log(1 + \alpha \delta)$$

Let $\xi(\alpha) = \frac{\alpha(\rho + \delta)}{1 + \alpha \delta} - \log(1 + \alpha \delta)$. We find that

$$\frac{d}{d\alpha} \xi(\alpha) = \frac{\rho - \delta^2 \alpha}{(1 + \alpha \delta)^2}$$

and that $\xi(\alpha)$ is maximized at $\alpha = \frac{\rho}{\delta^2}$. Substituting this value of α into our inequality, we obtain

$$(4.1) \quad f(x + s) \leq f(x) - \omega(\eta)$$

□

Since $\omega(\eta) = \eta - \log(1 + \eta)$ is positive for all $\eta > 0$, it follows that if $\limsup_k \eta_k > 0$, then $f(x_k) \rightarrow -\infty$. This simple fact will be crucial in our convergence analysis.

Lemma 4.2. *If f is bounded below, then $\eta_k \rightarrow 0$.*

It will also be useful to relate the step sizes to η . Observe that

$$(4.2) \quad t = \frac{\alpha}{1 + \alpha\delta} = \frac{\eta}{\delta(1 + \eta)}$$

and thus

$$(4.3) \quad t\rho = \frac{\eta^2}{1 + \eta}$$

5. SCALED GRADIENT METHODS

We first consider the class of methods where the matrices H_k are positive definite and uniformly bounded above and below. That is, there exist positive constants λ, Λ such that for every $k \geq 0$,

$$(5.1) \quad \lambda I \preceq H_k \preceq \Lambda I$$

The convergence analysis is particularly straightforward for these methods.

We assume that f is self-concordant, bounded below, and the Hessian of f is bounded above on the level set $\Omega = \{x : f(x) \leq f(x_0)\}$. Let M be a constant such that $G(x) \preceq MI$ for all $x \in \Omega$.

Theorem 5.1. *Consider any adaptive method for which the matrices H_k satisfy eq. (5.1). Then the method converges in the sense that $\lim_k \|g_k\| = 0$.*

Proof. Observe that

$$(5.2) \quad \eta_k = \frac{g_k^T H_k g_k}{\sqrt{g_k^T H_k G(x_k) H_k g_k}} \geq \frac{\lambda}{\Lambda \sqrt{M}} \|g_k\|$$

By Lemma 4.2, $\eta_k \rightarrow 0$. Therefore $\|g_k\| \rightarrow 0$. □

If in addition, f is strongly convex with $mI \preceq G(x)$ for $m > 0$, then an adaptive method satisfying eq. (5.1) is globally R -linearly convergent. The proof uses the following lemma, which is well known (for instance, see [2, 5]).

Lemma 5.2 (Lemma 5.3, [5]). *If f is strongly convex with $mI \preceq G(x)$, and x_* is the unique minimizer of f , then $\|g(x)\|^2 \geq 2m(f(x) - f(x_*))$.*

Theorem 5.3. *If f is self-concordant and strongly convex (so there exist constants $0 < m \leq M$ such that $mI \preceq G(x) \preceq MI$ for all $x \in \Omega$), then an adaptive method for which the matrices H_k satisfy eq. (5.1) is globally R -linearly convergent. That is, there exists a constant $\gamma < 1$ such that $f(x_{k+1}) - f(x_*) \leq \gamma(f(x_k) - f(x_*))$ for all k .*

Proof. Since $\eta_k \rightarrow 0$ by Lemma 4.2, the sequence $\{\eta_k\}_{k=0}^\infty$ is bounded. Let $\Gamma = \sup_k \eta_k < \infty$, and let $c = \frac{1}{2(1+\Gamma)}$. Observe that $\omega(z) \geq cz^2$ for $0 \leq z \leq \Gamma$, as $\omega(0) = 0$ and $\frac{d}{dz}(\omega(z) - cz^2) = \frac{z(1-2c-2cz)}{1+z}$ which is non-negative for $0 \leq z \leq \Gamma$. Hence, since $\eta_k \leq \Gamma$ for all k , we have

$$\begin{aligned} f(x_{k+1}) - f(x_*) &\leq f(x_k) - f(x_*) - \omega(\eta_k) \leq f(x_k) - f(x_*) - c\eta_k^2 \\ &\leq f(x_k) - f(x_*) - \frac{c\lambda^2}{\Lambda^2 M} \|g(x_k)\|^2 \\ &\leq \left(1 - \frac{\lambda^2 m}{\Lambda^2(1+\Gamma)M}\right) (f(x_k) - f(x_*)) \end{aligned}$$

where the first line follows from inequality (3.1), the second from inequality (5.2), and the third from Lemma 5.2. Taking $\gamma = 1 - \frac{\lambda^2 m}{\Lambda^2(1+\Gamma)M}$, we obtain the desired R -linear convergence. \square

5.1. Adaptive Gradient Descent. When $H_k = I$, the method corresponds to gradient descent with adaptive step sizes incorporating second-order information. This strategy for selecting step sizes may have several advantages in practice. Using second-order information allows a better local model of the objective function. The classical analysis of gradient descent with a fixed step size also generally requires a sufficiently small step size in order to guarantee convergence. This step size is a function of the Lipschitz constant of the gradient $g(x)$, which is either unknown or impractical to compute. The step size needed to ensure convergence in theory is also often impractically tiny, leading to slow convergence in practice. For the class of self-concordant functions, an adaptive step size can be easily computed without knowledge of any constants, and still provides a theoretical guarantee of convergence, which is a significant advantage.

A proximal gradient descent method with adaptive step sizes was studied by Tran-Dinh et al. [18], who proved the method to be globally convergent for self-concordant functions, and locally R -linearly convergent for strongly convex self-concordant functions. However, our convergence analysis above employs different techniques from those in [18], and in particular, we obtain the following theorem as an immediate corollary of Theorem 5.1 and Theorem 5.3:

Theorem 5.4. *Suppose that f is self-concordant, bounded below, and $G(x) \preceq MI$ on the level set $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$. Then the adaptive gradient descent method converges in the sense that $\lim_{k \rightarrow \infty} \|g_k\| = 0$. Furthermore, if f is strongly convex on Ω , then the adaptive gradient descent method is globally R -linearly convergent.*

5.2. Adaptive L-BFGS. The limited-memory BFGS algorithm (L-BFGS, [9]) stores a fixed number of previous *curvature pairs* (s_k, y_k) , where $y_k = g_{k+1} - g_k$, and computes $d_k = -H_k g_k$ from the curvature pairs using a two-loop recursion [13]. It is well-known that L-BFGS satisfies eq. (5.1). In [6], the following bounds are obtained.

Theorem 5.5 (Lemma 1, [6]). *Let ℓ be the number of curvature pairs stored by the L-BFGS method. Then the matrices H_k satisfy*

$$\lambda I \preceq H_k \preceq \Lambda I$$

where $\lambda = (1 + \ell M)^{-1}$ and $\Lambda = (1 + \sqrt{\kappa})^{2\ell} \left(1 + \frac{1}{m(2\sqrt{\kappa} + \kappa)}\right)$ for $\kappa = M/m$.

Hence, it follows immediately from Theorem 5.1 and Theorem 5.3 that:

Theorem 5.6. *Suppose that f is self-concordant, bounded below, and $G(x) \preceq MI$ on the level set $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$. Then the adaptive L-BFGS method converges in the sense that $\lim_{k \rightarrow \infty} \|g_k\| = 0$. Furthermore, if f is strongly convex on Ω , then the adaptive L-BFGS method is globally R-linearly convergent.*

We note that, as with gradient descent, the classical analysis of L-BFGS requires either that inexact Armijo-Wolfe line search is performed, or that a sufficiently small fixed step size, depending on the Lipschitz constant of $g(x)$, is used.

6. ADAPTIVE BFGS

If H_k is chosen to approximate $(\nabla^2 f(x_k))^{-1}$, then we obtain quasi-Newton methods with adaptive step sizes. In particular, we may iteratively update H_k using the BFGS update formula, which we briefly describe. Let $B_k = H_k^{-1}$, and let $y_k = g_{k+1} - g_k$. The BFGS update sets B_{k+1} to be the nearest matrix to B_k (in a variable metric) satisfying the *secant equation* $B_{k+1}s_k = y_k$. It is well known that H_{k+1} has the following expression in terms of H_k, s_k and y_k :

$$(6.1) \quad H_{k+1} = \frac{s_k s_k^T}{y_k^T s_k} + \left(I - \frac{s_k y_k^T}{y_k^T s_k} \right) H_k \left(I - \frac{y_k s_k^T}{y_k^T s_k} \right)$$

6.1. Superlinear Convergence of Adaptive BFGS. The convergence analysis of the classical BFGS method [15, 2] assumes that the method uses inexact line searches satisfying the *Armijo-Wolfe* conditions: for constants $\alpha, \beta \in (0, 1)$ with $\alpha < \frac{1}{2}$ and $\alpha < \beta$, the step size λ_k should satisfy

$$(6.2) \quad f(x_{k+1}) \leq f(x_k) + \alpha \lambda_k g_k^T d_k$$

and

$$(6.3) \quad g(x_{k+1})^T d_k \geq \beta g(x_k)^T d_k$$

Powell [15] proves the following global convergence theorem for BFGS.

Theorem 6.1 (Lemma 1, [15]). *If the BFGS algorithm with Armijo-Wolfe inexact line search is applied to a convex function $f(x)$ that is bounded below, if x_0 is any starting vector and B_0 is any positive definite matrix, and if the Hessian $G(x)$ satisfies $G(x) \preceq MI$ for all x in the level set $\Omega = \{x : f(x) \leq f(x_0)\}$, then the limit*

$$(6.4) \quad \liminf_{k \rightarrow \infty} \|g_k\| = 0$$

is obtained.

In our setting, f is a self-concordant function that is bounded below and satisfies $G(x) \preceq MI$. In order to prove that adaptive BFGS is convergent in the sense of the limit (6.4), it suffices to show that the adaptive step sizes t_k eventually satisfy the Armijo-Wolfe conditions. Specifically, we prove the following two theorems that apply to *every* adaptive method.

Theorem 6.2. *The curvature-adaptive step size t_k satisfies the Armijo condition for any $\alpha < \frac{1}{2}$.*

Proof. Let $\alpha < \frac{1}{2}$. We aim to prove that $f(x_{k+1}) \leq f(x_k) + \alpha t_k g_k^T d_k$. By Lemma 4.1, $f(x_{k+1}) \leq f(x_k) - \omega(\eta_k)$. Therefore, it suffices to prove that

$$\omega(\eta_k) \geq -\frac{1}{2} t_k g_k^T d_k$$

For brevity, we omit the index k . Notice that $-tg^T d = tg^T Hg = t\rho$. By eq. (4.3), $t\rho = \frac{\eta^2}{1+\eta}$. Therefore, we must prove that for $\eta \geq 0$,

$$\eta - \log(1 + \eta) \geq \frac{1}{2} \frac{\eta^2}{1 + \eta}$$

Define $\zeta(z) = z - \log(1 + z) - \frac{1}{2} \frac{z^2}{1+z}$. Observe that $\zeta(0) = 0$ and

$$\frac{d}{dz} \zeta(z) = 1 - \frac{1}{1+z} - \frac{1}{2} \frac{z^2 + 2z}{(1+z)^2} = \frac{1}{2} \frac{z^2}{(1+z)^2}$$

Since $\frac{d}{dz} \zeta(z) \geq 0$ for all $z \geq 0$, we conclude that $\omega(\eta) \geq \frac{1}{2} \frac{\eta^2}{1+\eta}$ for all $\eta \geq 0$. This completes the proof. \square

Theorem 6.3. *Let $\beta < 1$. There exists an index k_0 such that for all $k \geq k_0$, the curvature-adaptive step size t_k satisfies the Wolfe condition.*

Proof. We aim to prove that $g(x_{k+1})^T d_k \geq \beta g(x_k)^T d_k$. This is equivalent to $g(x_k + t_k d_k)^T d_k - g(x_k)^T d_k \geq -(1 - \beta) g(x_k)^T d_k$. Observe that $-g(x_k)^T d_k = g(x_k)^T H_k g(x_k) = \rho_k$. By inequality (3.2) with $h = d_k$, we have

$$g(x_k + t_k d_k)^T d_k - g(x_k)^T d_k \geq \frac{\delta_k^2 t_k}{1 + \delta_k t_k} \rho_k$$

We omit the index k for brevity. Thus, it suffices to prove that

$$\frac{\delta^2 t}{1 + \delta t} \geq (1 - \beta) \rho$$

However,

$$\frac{\delta^2 t}{1 + \delta t} = \frac{1}{1 + 2\eta} \rho$$

Since $\eta \rightarrow 0$, there exists some k_0 such that $\frac{1}{1+2\eta_k} \geq 1 - \beta$ for all $k \geq k_0$. \square

We can then immediately apply Theorem 6.1 to deduce that adaptive BFGS is convergent.

Theorem 6.4. *Let f be self-concordant, bounded below, and assume the Hessian satisfies $G(x) \preceq MI$ for all $x \in \Omega$. Then for the adaptive BFGS method, $\liminf_{k \rightarrow \infty} \|g_k\| = 0$.*

It is well known that if the objective function f is strongly convex, then the classical BFGS method converges Q -superlinearly. Let us now assume that f is strongly convex, so there exists a constant $m > 0$ with $mI \preceq G(x)$ for all $x \in \Omega$. Let x_* denote the unique minimizer of f .

Theorem 6.5 (Lemma 4, [15]). *Let f be strongly convex, and let $\{x_k\}_{k=0}^\infty$ be the sequence of iterates generated by the BFGS method with inexact Armijo-Wolfe line searches. Then $\sum_{k=0}^\infty \|x_k - x_*\| < \infty$.*

Since the adaptive step size t_k eventually satisfies the Armijo-Wolfe conditions, the same holds for BFGS with adaptive step sizes.

Theorem 6.6. *Let f be self-concordant and strongly convex. The sequence of iterates $\{x_k\}_{k=0}^\infty$ produced by adaptive BFGS satisfies $\sum_{k=0}^\infty \|x_k - x_*\| < \infty$.*

In the proof of superlinear convergence for the classical BFGS method, it is assumed that the Hessian $G(x)$ is Lipschitz continuous. However, it is unnecessary to make this assumption in our setting, as $G(x)$ is necessarily Lipschitz when f is self-concordant and $G(x)$ is bounded above.

Theorem 6.7. *If f is self-concordant and satisfies $G(x) \preceq MI$ for all $x \in \Omega$, then $G(x)$ is Lipschitz continuous on Ω .*

Proof. Let $x, y \in \Omega$, and let $e = x - y$. Let $v \in \mathbb{R}^n$ be any unit vector. By Taylor's Theorem, we have

$$v^T G(x)v = v^T G(y)v + \int_0^1 \nabla^3 f(y + \tau e)[v, v, e] d\tau$$

Hence, by Theorem 3.1,

$$\begin{aligned} |v^T (G(x) - G(y))v| &\leq \int_0^1 |\nabla^3 f(y + \tau e)[v, v, e]| d\tau \\ &\leq 2 \int_0^1 v^T G(y + \tau e)v \sqrt{e^T G(y + \tau e)e} d\tau \\ &\leq 2 \int_0^1 M^{3/2} \|e\| d\tau = 2M^{3/2} \|x - y\| \end{aligned}$$

Therefore, the eigenvalues of $G(x) - G(y)$ are bounded in norm by $2M^{3/2} \|x - y\|$. It follows that $\|G(x) - G(y)\| \leq 2M^{3/2} \|x - y\|$, so $G(x)$ is Lipschitz continuous. \square

It is well known that the BFGS method is invariant under an affine change of coordinates, so we may assume without loss of generality that $G(x_*) = I$. This corresponds to considering the function $\tilde{f}(\tilde{x}) = f(G(x_*)^{-1/2}\tilde{x})$ and performing a change of coordinates $\tilde{x} = G(x_*)^{1/2}x$. By Theorem 4.1.2 of [11], the function \tilde{f} is also self-concordant.

To complete the proof of superlinear convergence, we use results established by Dennis and Moré [4] and Griewank and Toint [7]. In §4 of [7], Griewank and Toint prove that, given Theorem 6.6 and the bound in Theorem 6.7, the following limit holds:

$$(6.5) \quad \lim_{k \rightarrow \infty} \frac{\|(B_k - I)d_k\|}{\|d_k\|} = 0$$

Furthermore, the argument in [7] shows that both $\{\|B_k\|\}_{k=0}^\infty$ and $\{\|H_k\|\}_{k=0}^\infty$ are bounded. Writing $B_k d_k = -g_k$ and $-d_k = H_k g_k$, we have an equivalent limit

$$(6.6) \quad \lim_{k \rightarrow \infty} \frac{\|H_k g_k - g_k\|}{\|g_k\|} = 0$$

This enables us to prove that the adaptive step sizes t_k converge to 1, which is necessary for superlinear convergence.

Theorem 6.8. *The curvature-adaptive step sizes t_k in the adaptive BFGS method converge to 1.*

Proof. We omit the index k for brevity, and define $u = Hg - g$. Since $t = \frac{\alpha}{1+\alpha\delta}$, it suffices to show that α converges to 1.

$$\begin{aligned}\alpha &= \frac{\rho}{\delta^2} = \frac{g^T Hg}{g^T HGHg} \\ &= \frac{g^T g + g^T u}{g^T Gg + 2g^T Gu + u^T Gu} \\ &= \frac{1 + \frac{g^T u}{g^T g}}{\frac{g^T Gg}{g^T g} + 2\frac{g^T Gu}{g^T g} + \frac{u^T Gu}{g^T g}}\end{aligned}$$

The Cauchy-Schwarz inequality, the fact that $G(x) \preceq MI$, and the limit (6.6) imply that $\frac{g^T u}{g^T g}$, $\frac{g^T Gu}{g^T g}$, $\frac{u^T Gu}{g^T g}$ converge to 0. Since $G = G(x_k)$ and $x_k \rightarrow x_*$, we have $G \rightarrow I$, and therefore $\frac{g^T Gg}{g^T g} \rightarrow 1$. It follows that α , and therefore t , converges to 1. \square

We now make a slight modification to the Dennis-Moré characterization of superlinear convergence. Write

$$\begin{aligned}\frac{\|(B_k - I)s_k\|}{\|s_k\|} &= \frac{\|t_k g_{k+1} - t_k g_k - G(x_*)s_k + t_k g_{k+1}\|}{\|s_k\|} \\ &\geq t_k \frac{\|g_{k+1}\|}{\|s_k\|} - \frac{\|t_k g_{k+1} - t_k g_k - t_k G(x_*)s_k + (1 - t_k)G(x_*)s_k\|}{\|s_k\|} \\ &\geq t_k \frac{\|g_{k+1}\|}{\|s_k\|} - \frac{t_k \|\int_0^1 (G(x_k + \tau s_k) - G(x_*))s_k d\tau\|}{\|s_k\|} - |1 - t_k| \frac{\|G(x_*)s_k\|}{\|s_k\|}\end{aligned}$$

Both of the latter terms converge to 0, and eq. (6.5) implies that $\frac{\|(B_k - I)s_k\|}{\|s_k\|}$ converges to 0, so we deduce that $\frac{\|g_{k+1}\|}{\|s_k\|}$ converges to 0.

However, since f is strongly convex, $\|g(x)\| \geq m\|x - x_*\|$. Hence, we find that

$$\frac{\|g_{k+1}\|}{\|s_k\|} \geq \frac{m\|x_{k+1} - x_*\|}{\|x_{k+1} - x_*\| + \|x_k - x_*\|},$$

which implies that $\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \rightarrow 0$. Thus, we have the following:

Theorem 6.9. *Suppose that f is self-concordant, and strongly convex on $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$. Then the adaptive BFGS method converges Q -superlinearly.*

By the same reasoning, the results in [2] and [7] imply that these convergence theorems also hold for the adaptive versions of the quasi-Newton methods in *Broyden's convex class*, with the exception of the DFP method. The adaptive versions of the Block BFGS methods proposed in [5] can also be shown to be Q -superlinearly convergent.

6.2. Hybrid Step Selection. Observe that when $\alpha \approx 1$, we have

$$t = \frac{\alpha}{1 + \alpha\delta} \approx \frac{1}{1 + \delta}$$

Consequently, t is small when δ is large, and a small step $t_k d_k$ is likely to result in t_{k+1} also being small. Thus, when the initial δ is large, a method using adaptive step sizes may produce a long succession of small steps. This suggests the following heuristic for selecting step sizes:

- (1) Select an array T_k of candidate step sizes for t_k .
- (2) At step k , test the elements of T_k in order until a candidate step size is found which satisfies the Armijo condition (6.2).
- (3) If no element of T_k satisfies the Armijo condition, then set t_k to be the adaptive step size.

For instance, in our numerical experiments reported in Section 7, we take T_k to be $(1, \frac{1}{4}, \frac{1}{16})$ for all k . This allows the method to take steps of size $t_k = 1$ when 1 satisfies the Armijo condition, which is desirable for speeding up superlinear convergence.

We refer to this scheme as *hybrid step selection*. For a proper choice of T_k , hybrid step selection avoids the disadvantage of exclusively using adaptive step sizes, where the step size may be small for many iterations. It will also generally be more efficient to compute than a full line search, since no more than $|T_k|$ candidate step sizes are tested before switching to the adaptive step size.

7. NUMERICAL EXPERIMENTS

To compare our adaptive methods to classical algorithms, we solve several binary classification problems using *logistic regression*. In these problems, the objective function to be minimized has the form

$$(7.1) \quad L(w) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i x_i^T w)) + \frac{1}{2m} \|w\|_2^2$$

where the training data $\{(x_1, y_1), \dots, (x_m, y_m)\}$ consists of feature vectors $x_i \in \mathbb{R}^n$ and classifications $y_i \in \{-1, +1\}$. Zhang and Xiao [19] showed that logistic regression is self-concordant.

Theorem 7.1 (Lemma 1, [19]). *Let $B = \max_i \|x_i\|$. The scaled function $\frac{B^2 m}{4} L(w)$ is standard self-concordant.*

In our tests, we compared seven algorithms:

- (1) BFGS with adaptive step sizes (BFGS-A).
- (2) BFGS with Armijo-Wolfe line search (BFGS-LS).
- (3) BFGS with hybrid step selection (BFGS-H), using $T_k = (1, \frac{1}{4}, \frac{1}{16})$.
- (4) L-BFGS with adaptive step sizes (LBFGS-A), using the past $\ell = \min\{\frac{n}{2}, 100\}$ curvature pairs.
- (5) L-BFGS with Armijo-Wolfe line search (LBFGS-LS).
- (6) Gradient descent with adaptive step sizes (GD-A).
- (7) Gradient descent with Armijo-Wolfe line search (GD-LS).

An initial Hessian approximation H_0 must be provided for the BFGS and L-BFGS methods. It is easy, but not necessarily effective, to simply take $H_0 = I$. Another common strategy for initializing H_0 , described in [14], that is often quite effective, is to take $H_0 = I$ on the first step, and then, before performing the first BFGS update (6.1), scale H_0 :

$$(7.2) \quad H_0 \leftarrow \frac{y_0^T s_0}{y_0^T y_0} I$$

It is easy to verify that the scaling factor $y_0^T s_0 / y_0^T y_0$ lies between the smallest and largest eigenvalues of the inverse of the average Hessian $\bar{G} = \int_0^1 G(x_0 + \tau s_0) d\tau$ along the initial step.

Data set	n	m
<code>covtype.libsvm.binary.scale</code>	55	581012
<code>ijcnn1.tr</code>	23	35000
<code>leu</code>	7130	38
<code>rcv1_train.binary</code>	47237	20242
<code>real-sim</code>	20959	72309
<code>w8a</code>	301	49749

TABLE 1. Data sets used in Section 7

Similarly, for the L-BFGS method, the initial matrix used at step $k + 1$ in the two-loop recursion is chosen as:

$$H_0 \leftarrow \frac{y_k^T s_k}{y_k^T y_k} I$$

We refer to this as *identity scaling*.

The line search used the `WolfeLineSearch` routine from the `minFunc` software package [17]. The Armijo-Wolfe parameters were $\alpha = 0.1, \beta = 0.75$, and the line search was configured to use an initial step size $t = 1$ and perform quadratic interpolation (`LS_interp = 1`, `LS_multi = 0`).

We chose six data sets from LIBSVM [3] with a variety of dimensions, which are listed in Table 1. We plot the progress of each algorithm as a function of CPU time used. The progress is measured by the *log gap* $\log_{10}(f(w) - f(w_*))$, where w_* is a pre-computed optimal solution. The starting point was always set to $w_0 = 0$. All algorithms were terminated when either the gradient reached the threshold $\|g(x)\| < 10^{-7}$, or after 480 seconds of CPU time.

Our algorithms were implemented in Matlab 2015a and run on an Intel i7-3630QM processor. While the CPU time is clearly platform-dependent, we sought to minimize implementation differences between the algorithms to make the test results as comparable as possible.

In Figure 1, we plot the results for the data sets `covtype.libsvm.binary.scale`, `ijcnn1.tr`, and `w8a`. On these problems, we implemented BFGS with a dense Hessian; that is, the matrices H_k were stored explicitly and updated using the formula (6.1).

In Figure 2, we plot the results for the data sets `leu`, `rcv1_train.binary`, and `real-sim`. These problems had a large number of variables ($n > 7000$), which made it infeasible to store H_k explicitly. On these problems, BFGS was also implemented using the two-loop recursion, with unlimited memory. If the number of iterations exceeds roughly $n/4$, then this approach would in fact require more memory than storing H_k explicitly. However, this never occurred in our tests, as shown in Table 2.

In our tests, we found that BFGS-A required more time than BFGS-LS. Although the cost of a single step was significantly lower for BFGS-A than BFGS-LS, BFGS-A often took numerous small steps in succession, making very slow progress. This situation was exactly our motivation for devising the hybrid step selection described in Section 6.2, and unfortunately, appears to occur often. However, BFGS-H achieved the same speed as BFGS-LS with $T = (1, \frac{1}{4}, \frac{1}{16})$, which suggests that always trying $t = 1$ first is an excellent heuristic.

Curiously, L-BFGS was far more effective on the problems with large n (Figure 2) than on those with small n (Figure 1). Both LBFGS-A and LBFGS-LS were ineffective on the problems with small n , which suggests that the problem lies with the step directions

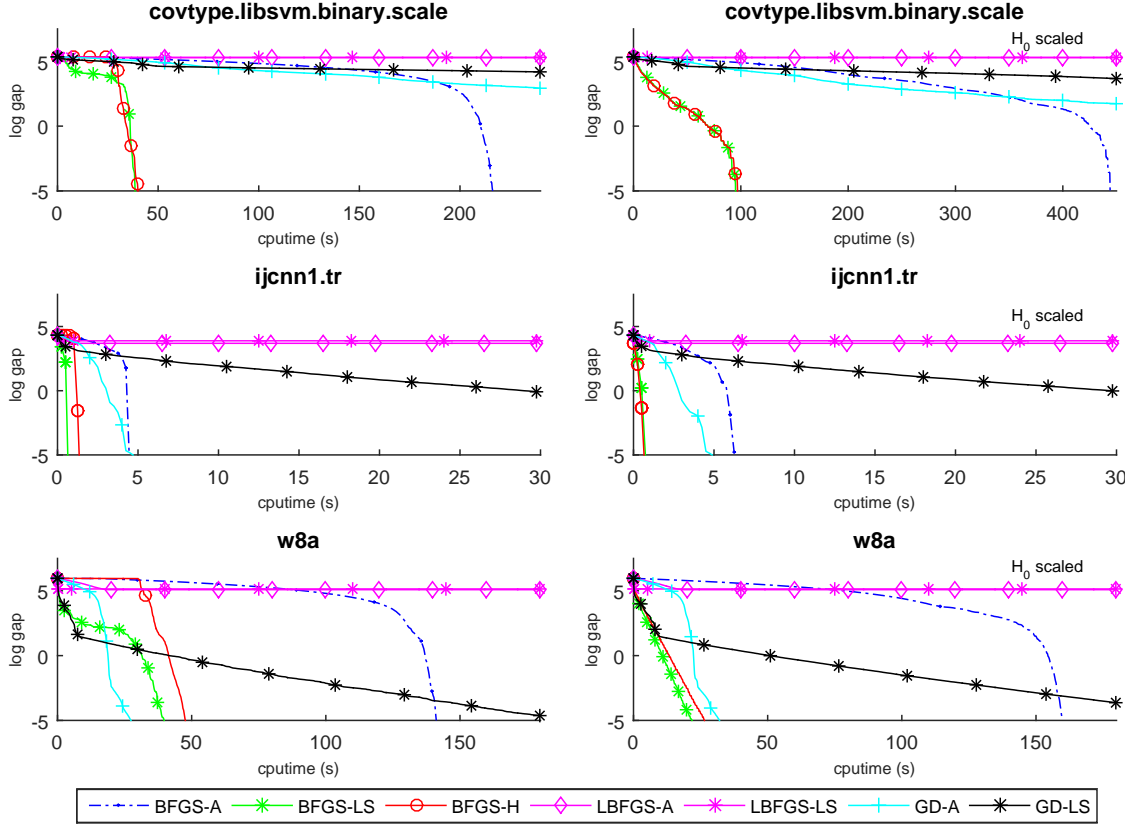


FIGURE 1. Experiments on problems with small n . The log gap is defined as $\log_{10}(f(w) - f(w_*))$. The loss functions are scaled to be standard self-concordant. All BFGS and L-BFGS plots on the left take $H_0 = I$, and those on the right use identity scaling.

Data set	Identity Scaling	n	Number of iterations		
			BFGS-A	BFGS-LS	BFGS-H
leu	No	7130	1197	97	314
leu	Yes	7130	908	179	263
rcv1_train.binary	No	47237	161	31	36
rcv1_train.binary	Yes	47237	284	206	229
real-sim	No	20959	356	42	55
real-sim	Yes	20959	592	251	317

TABLE 2. The number of iterations until convergence of the BFGS methods on problems with large n .

computed by L-BFGS, rather than the step sizes. Identity scaling was also beneficial for L-BFGS on problems with large n , substantially reducing the convergence time in some cases. On the other hand, identity scaling appeared to be detrimental for the BFGS methods *only* on the data set **leu**. The data set **leu** appears to be quite different from the other problems tested.

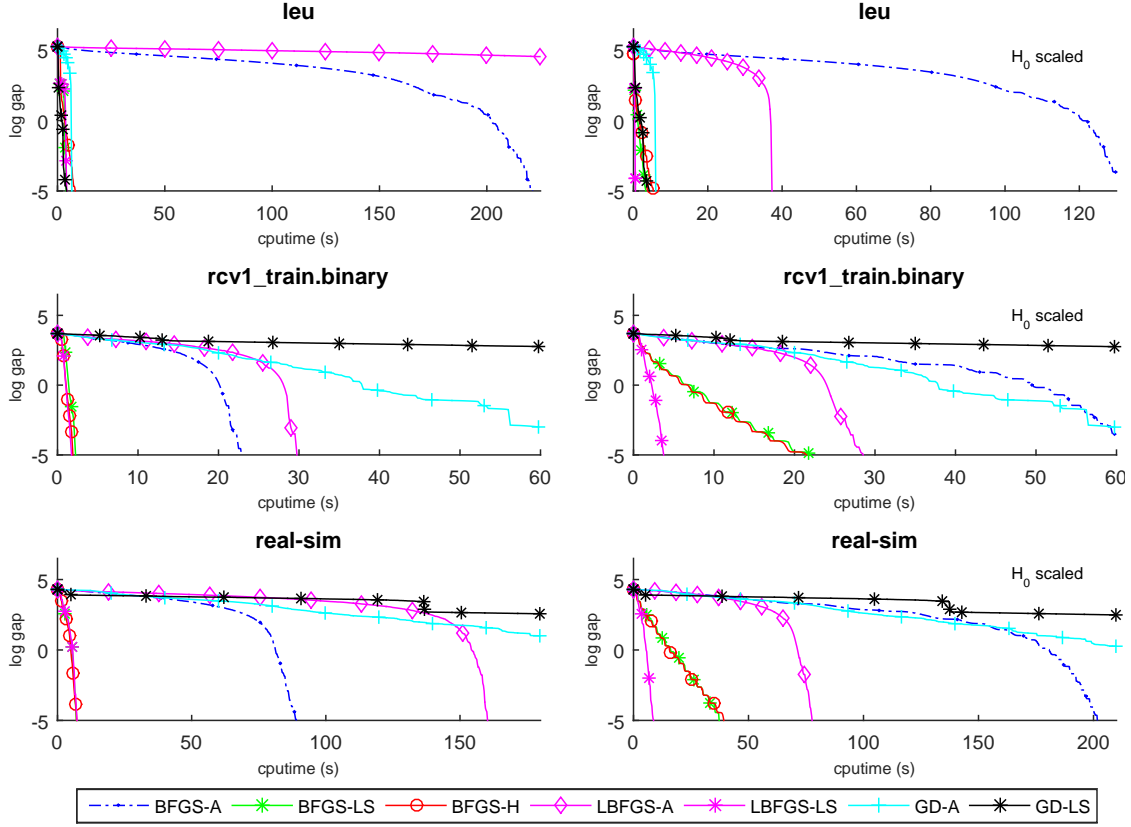


FIGURE 2. Experiments on problems with large n . The log gap is defined as $\log_{10}(f(w) - f(w_*))$. The loss functions are scaled to be standard self-concordant. All BFGS and L-BFGS plots on the left take $H_0 = I$, and those on the right use identity scaling.

The number of training samples for **leu** was $m = 38$, while for all other problems, m was at least 20,000. Moreover, gradient descent with Armijo-Wolfe line search (GD-LS) was among the fastest methods on **leu**, while on the other test problems it was significantly outperformed by BFGS.

On the other hand, GD-A was surprisingly effective, often outperforming GD-LS. This is very encouraging, as adaptive gradient descent can readily be extended to stochastic optimization (see Section 8), which has important applications.

8. FURTHER DIRECTIONS

The adaptive step size can readily be extended to *stochastic* optimization methods. Consider a problem of the form

$$(8.1) \quad L(w) = \frac{1}{m} \sum_{i=1}^m f_i(w) + h(w)$$

If m is extremely large, as is often the case in machine learning, simply evaluating $L(w)$ is an expensive operation, and line search is entirely impractical. To solve problems of the form (8.1), stochastic algorithms such as Stochastic Gradient Descent (SGD, [1]) select a

random subset S_k of $\{f_1, \dots, f_m\}$ at step k and take a single step using the gradient for the subsampled problem

$$(8.2) \quad L^{(S_k)}(w) = \frac{1}{|S_k|} \sum_{f_i \in S_k} f_i(w) + h(w)$$

as an approximation to the gradient of the loss function (8.1). In variance-reduced versions of SGD such as SVRG [8], it is common to use a constant step size, determined through experimentation.

In fact, this setting was the motivation for our interest in adaptive methods. Our goal was to devise a method for solving (8.1) by computing the subsampled gradient of (8.2) at each iteration, and proceeding without using line search or ad hoc empirical selection to obtain the step sizes. The curvature-adaptive step size can readily be computed for the subsampled problem (8.2), so long as we have access to Hessian-vector products. Furthermore, the curvature-adaptive step size incorporates second-order information, which is currently not exploited by most stochastic algorithms.

We have obtained several preliminary results for stochastic versions of the adaptive methods. For adaptive gradient descent (Section 5.1), we can show that the stochastic adaptive gradient descent method returns an ϵ -optimal solution in expectation after $O(\log(\frac{1}{\epsilon}))$ iterations if S_k is chosen to be a subset of size $O(\frac{1}{\epsilon})$.

It would be of great interest to show that a stochastic version of adaptive BFGS (Section 6) converges superlinearly. We are only aware of one specialized algorithm for minimizing problems with the finite sum structure (8.1) which is provably superlinear - the Newton Incremental Method (NIM) of Rodomanov and Kropotov [16]. However, incremental-type methods such as NIM have memory requirements of the order $O(m)$, which is often substantial. We are hopeful that adaptive BFGS will lead to new results on superlinear convergence in stochastic methods.

REFERENCES

- [1] L. BOTTOU, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMSTAT'2010, Physica-Verlag HD, pp. 177–186.
- [2] R. H. BYRD, J. NOCEDAL, AND Y.-X. YUAN, *Global convergence of a class of quasi-Newton methods on convex problems*, Siam. J. Numer. Anal., (1987), pp. 1171–1190.
- [3] C.-C. CHANG AND C.-J. LIN, *LIBSVM: A library for support vector machines*, ACM Trans. Intell. Syst. Technol., 2 (2011).
- [4] J. E. DENNIS JR. AND J. J. MORÉ, *Characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.
- [5] W. GAO AND D. GOLDFARB, *Block BFGS methods*, in review. arXiv:1609.00318, (2016).
- [6] R. GOWER, D. GOLDFARB, AND P. RICHTÁRIK, *Stochastic block BFGS : Squeezing more curvature out of data*, in JMLR: Workshop and Conference Proceedings, vol. 48, 2016.
- [7] A. GRIEWANK AND P. L. TOINT, *Local convergence analysis for partitioned quasi-Newton updates*, Numer. Math., 39 (1982), pp. 429–448.
- [8] R. JOHNSON AND T. ZHANG, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems, 2013, pp. 315–323.
- [9] D. LIU AND J. NOCEDAL, *On the limited memory BFGS method for large scale optimization*, Math. Prog., 45 (1989), pp. 503–528.
- [10] Z. LU, *Randomized block proximal damped newton methods for composite self-concordant minimization*, in review. arXiv:1607.00101, (2016).
- [11] Y. NESTEROV, *Introductory Lectures on Convex Optimization*, Springer Science+Business Media, New York, 2004.

- [12] Y. NESTEROV AND A. NEMIROVSKI, *Interior-point polynomial algorithms in convex programming*, Society for Industrial and Applied Mathematics, Philadelphia, 1994.
- [13] J. NOCEDAL, *Updating quasi-Newton matrices with limited storage*, Math. Comp., 35 (1980), pp. 773–782.
- [14] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer Science+Business Media, New York, 2nd ed., 2006.
- [15] M. J. D. POWELL, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming, R. Cottle and C. Lemke, eds., vol. IX, SIAM-AMS Proceedings, 1976.
- [16] A. RODOMANOV AND D. KROPOTOV, *A superlinearly-convergent proximal newton-type method for the optimization of finite sums*, in Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 2597–2605.
- [17] M. SCHMIDT, *minFunc: unconstrained differentiable multivariate optimization in matlab*, <http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>, (2005).
- [18] Q. TRAN-DINH, A. KYRILLIDIS, AND V. CEVHER, *Composite self-concordant minimization*, Journal of Machine Learning Research, 16 (2015), pp. 371–416.
- [19] Y. ZHANG AND L. XIAO, *Disco: Distributed optimization for self-concordant empirical loss*, in JMLR: Workshop and Conference Proceedings, vol. 32, 2015, pp. 362–370.